International Conference "Water, Environment, Energy and Society" (WEES–2009)
New Delhi, 12–16 January 2009

Rainfall Station Network Optimization Using Entropy Method

P.G. Jairaj[1] and A.R. Remya
Department of Civil Engineering
College of Engineering, Trivandrum, Kerala - 695016, INDIA
E-mail: [1]jairaj_pg@yahoo.com

ABSTRACT: Data collection forms the basis for most of the analysis and decision making process in hydrological studies. An efficient data collection network is to be designed so as to minimize the expenditure in data collection, at the same time incorporating its spatial variability. The application of entropy concept which gives a quantitative measure of information is used to express the redundancy in networks, and to derive an optimum network. The present study deals with the methodology based on entropy concept for ranking the stations of a rainfall gauging station network. The application of the methodology to the monthly rainfall values of the rain gauge network stations of United Kingdom is illustrated. The significance of the assumed distribution to the station data series, in identifying the optimum data collection network is also analysed.

INTRODUCTION

Data regarding various parameters are needed for most of the analysis and design purposes in the field of water resources engineering. In order to achieve this, data collecting networks like stream flow networks, rainfall networks, water quality networks etc are used. Data collections involves large amount of money and manpower. Developing countries cannot afford to spend large amounts for data collection; and this necessitates the need for optimizing the data collecting networks. This will help to achieve economy without compromising on the quality of data.

Entropy is a quantitative measure of uncertainty in a system. (Singh, 2000), gives a concise and brief overview of the concepts of entropy and its field applications in hydrology water resources modeling studies. An entropy based method (Ozkul *et al.*, 2000), has been developed to rank the data collecting stations in a network based on the redundant values they produce. The method has been applied to some data collecting networks like water quality monitoring networks (Ozkul *et al.*, 2000), and stream flow networks (Sarlak and Sorman, 2006).

In the present study the entropy based methodology has been illustrated for ranking the stations in a rainfall measuring stations network. The procedure was applied to the monthly rainfall values of the rain gauge stations network in United Kingdom. Two types of distributions viz. normal and lognormal have been assumed to the monthly station rainfall data, and its sensitivity to the optimized network is studied. A brief review about the concepts and forms of entropy, methodology adopted and the application to a real field case study and the sensitivity of the results to the input data series distributions are explained in the subsequent sections.

ENTROPY CONCEPT

Entropy may be termed as a measure of uncertainty associated with a process. The entropy theory is comprised of three main parts: Shannon Entropy, Principle of Maximum Entropy and Principle of Maximum Cross Entropy. Depending on the situation and type of variable(s) whose uncertainty is to be determined there are different types of entropies namely Marginal Entropy, Conditional Entropy, Trans-information etc.

Shannon Entropy

The probability distribution of events if known provides a certain amount of information. The uncertainty can be quantified with the entropy taking into account all different kinds of available information. Thus entropy is a measure of uncertainty represented by the probability distribution and is a measure of the chaos or of the lack of information about a system. If complete information is available, entropy is equal to zero; otherwise it is greater than zero.

Shannon defined a quantitative measure of the uncertainty associated with a probability distribution or the information content of the distributions in terms

[1]Conference speaker

of entropy called Shannon Entropy (Kapur and Kesavan, 1992), mathematically expressed as,

$$H(X) = -k\sum p_i \ln p_i \quad \ldots (1)$$

where $H(X)$ is the entropy corresponding to the random variable X, k is a constant which has the value equal to one when natural logarithm is taken. p_i represents the probability distribution corresponding to X.

Principle of Maximum Entropy

According to the Principle of Maximum Entropy (POME), when making inferences based on incomplete information, the probability distribution to be drawn must have the maximum entropy permitted by the available informations expressed in the form of constraints. According to the Shannon entropy as an information measure, the POME based distribution is favored over those with less entropy among those, which satisfy the given constraints. Intuitively, distributions of higher entropy represent more disorder, smoother, are more probable, are less predictable, or assume less. The POME based distribution (Singh, 2000), is maximally non-committal with regard to missing information. POME is expressed mathematically as,

$$\text{Maximize } H(X) = -k\sum p_i \ln p_i \quad \ldots (2)$$

subject to available information. Maximizing the entropy of the system will make the probability distribution as uniform as possible while satisfying the constraints.

Principle of Minimum Cross Entropy

According to Laplace's Principle of insufficient reason, all outcomes of an experiment should be considered equally likely unless there is information to contrary. Suppose we guess a probability distribution for a variable X as $Q = \{q_1, q_2, \ldots, q_n\}$ based on intuition or theory. This constitutes the prior information in terms of a prior distribution. To verify our guess, we take a set of observations, $X = \{x_1, x_2, \ldots, x_n\}$ and compute moments based on these observations. To derive the distribution $P = \{p_1, p_2, \ldots, p_n\}$ of X, we take all the given information and make the distribution as near to our intuition and experiences as possible. Thus the Principle of Minimum Cross Entropy (Singh, 2000), is expressed as,

$$\text{Minimize } D(P,Q) = \sum p_i \ln \frac{p_i}{q_i} \quad \ldots (3)$$

where $D(P, Q)$ is the cross entropy of the probability distribution P given the intuitive probability distribution Q.

Forms of Entropy

On the basis of the problem and the variables associated with it, the measure of entropy varies. There are four basic forms of entropy (Ozkul *et al.*, 2000), namely Marginal Entropy, Joint Entropy, Conditional Entropy and Transinformation. Each of these measures is used either individually or in combination as the situation demands.

Marginal Entropy

Marginal entropy represents the entropy of a single variable. It is expressed as,

$$H(X) = -k\sum_{i-1}^{i=n} p_i \ln p_i \quad \ldots (4)$$

where $H(X)$ is the entropy corresponding to the random variable X, k is the constant which has the value equal to one when natural logarithm is taken. p_i represents the probability distribution corresponding to X_i. n represents the number of elementary events with probabilities p_i.

Joint Entropy

It is the entropy corresponding to two or more variables. If the variables are considered as independent then their joint entropy is equal to the sum of their marginal entropies i.e., joint entropy $H(X, Y)$ is given as,

$$H(X, Y) = H(X) + H(Y) \quad \ldots (5)$$

If the variables are stochastically dependent their joint entropy is less than the total entropy given by Eqn. (5).

Conditional Entropy

It measures the entropy of a random variable Y, if one has already learned completely the value of the second random variable X. It is referred to as the entropy of Y conditioned on X, and is written as $H(Y|X)$. The conditional entropy is mathematically expressed as,

$$H(Y \mid X) = H(X,Y) - H(X) \quad \ldots (6)$$

Conditional entropy value becomes zero if the value of one variable is completely determined by the value of other variable. If the variables are independent the conditional entropy of Y given X becomes equal to $H(Y)$.

Transinformation

This is the form of entropy that measures the redundant or mutual information between variables. It

is described as the difference between the total entropy and the joint entropy of the variables. Consider two variables X and Y, the transinformation between these variables is given as,

$$T(X, Y) = H(Y) + H(X) - H(X, Y) \quad \ldots (7)$$

In this section the forms of entropy has been described considering two variables X and Y, but this can be extended to the multivariate case having M variables. Redundancy or repeated information in a network affects the efficiency of an information network. The various forms of entropy explained in this section helps to quantitatively represent the redundant information in the network. Measure of redundancy, forms the basis of the network optimization methodology explained in the next section.

METHODOLOGY

The main objective of optimizing an information network is to obtain maximum information with least number of stations. For this purpose it is necessary to identify stations producing redundant or repeated information. This is possible by quantitatively expressing the redundancy produced by each station. The redundancy can be expressed as a function of transinformation and joint entropy. The procedure developed for optimal information network based on entropy concept (Ozkul *et al.*, 2000) is discussed.

Consider an information network which collects data from more than one station. The approach here is to assess the reduction in the joint entropy of two or more variables due to the presence of stochastic dependence between them. This reduction is equivalent to the redundant information in the series at different sites. Thus the objective is to minimize transinformation by an appropriate choice of number and locations of monitoring stations. The combination of stations with least transinformation reflects the variability of the variable considered without producing redundant information. Thus the existing sampling sites can be sorted in the order of decreasing uncertainty or informativeness. In this ordered list, the first station is the one where the highest uncertainty about the variable occurs. The subsequent stations incorporated serve to reduce this uncertainty further so that the last station in the list brings least amount of information. Here it is possible to select a particular threshold transinformation value as the amount of redundant information to be permitted in the network such that sampling of the variable may be stopped at the stations that exceed the threshold.

The following procedure is applied for each variable to select the best combination of stations according to the entropy method.

1. It is assumed that there is M monitoring stations in the basin. The data series of the variable at each station is represented by X_m with the outcomes $x_{m,i}$ where $m(m = 1, 2, \ldots, M)$ denotes the station identification number and $i(i = 1, 2, \ldots, N)$ the time point along the sample. Here the sampling duration at all stations is considered to be equal. But the total number N of available data at each station can be different because there are often missing values or gaps within the data series.
2. Next, the type of the multivariate joint probability density function which best fits the distribution of $X_m(m = 1, 2, \ldots, .M)$ is selected. For a multivariate normal distribution the joint entropy of M stations, $H(X_1, X_2, \ldots, .X_m)$ can be calculated using the equation given below,

$$H(X) = \frac{M}{2}\ln 2\pi + \frac{1}{2}\ln|C| + \frac{M}{2} \quad \ldots (8)$$

where $|C|$ is the determinant of the coefficient matrix C. This joint entropy represents the total uncertainty about the particular variable in the network, which is to be reduced by sampling M monitoring stations.

3. The marginal entropy of the variable for each station X_m is calculated using the equation (8) by replacing the value of M by one. The station with the highest marginal entropy is denoted as the first priority station X_1; this is the location where the highest uncertainty occurs about the variable so that the highest information may be gained by making observations at this site. Note that the stations identification number 'm' is now being replaced by priority index j such that the m^{th} station X_m with the highest entropy is denoted as $X_j = X_1$.
4. Later, this station is coupled with every other station in the network to select that pair which gives the least transinformation. The station that fulfills this condition is marked as the second priority location $X_j = X_2$ such that,

$$\min\{H(X_m) - H(X_1 \mid X_2)\} = \min\{T(X_1, X_2)\} \ldots (9)$$

In the next step, the pair (X_1, X_2) is coupled with every other station in the network to select a triple with the least transinformation.

5. The same procedure is continued by successively considering combinations of 3, 4, 5, …., j stations

and selecting the combination that produces the least transinformation by satisfying the condition,

$$\min\{H(X_1,.....X_{j-1}) - H(X_1,....,X_{j-1} \mid X_j)\} = \min\{T(X_1....,X_{j-1}),X_j)\} \quad \ldots (10)$$

where X_1 is the first priority station and X_j is the station with the j^{th} priority. For multivariate normal probability density function, transinformation can also be determined by,

$$T\{(X_1,.....,X_{j-1}),X_j\} = -\frac{1}{2}\ln(1-R^2) \quad \ldots (11)$$

where R represents the multiple correlation coefficient. Accordingly, the above procedure assures the selection of a station X_j that has the least correlation with other stations in the network. In carrying out the above procedure, one can evaluate the results at each step by defining percentage of non transferred and transferred information among stations as,

$$t_j = \frac{H(X_1,.....,X_{j-1} \mid X_j)}{H(X_1,.....,X_{j-1})} \quad \ldots (12)$$

$$1-t_j = \frac{T((X_1,.....,X_{j-1}),X_j)}{H(X_1,.....,X_{j-1})} \quad \ldots (13)$$

Here, the designer may decide how much repeated information he wants to permit in the network. If he specifies this upper limit of redundant information as $(1 - t_j)^*$ percentage, he can select the combination of stations that produces this percentage as the one that must be included in the network. Stations that are added to the system after reaching $(1 - t_j)^*$ will increase the redundant information further so that one may decide to quit monitoring at such locations.

6. The evaluation explained in the last step can be made by defining k_j or the ratio of uncertainty explained by j number of stations to that explained by the total M number of stations in the network,

$$k_j = \frac{H(X_1,.....,X_{j-1} \mid X_j)}{H(X_1,.....,X_M)} \quad \ldots (14)$$

One may specify here an upper limit k_j^* as the percentage of uncertainty that is to be removed by the network. This upper limit is reached by a certain combination of stations; thus sampling sites that produce $k_j = k_j^*$ may be discontinued.

In the above procedure, the benefits for each combination of sampling sites are measured in terms of least transinformation or the highest conditional entropy produced by that combination. To select the best combinations of stations, it is sufficient to compare the costs and benefits represented by t_j or k_j. The above procedure helps to assess network configurations with respect to the existing stations. If new stations are to be added to the system, their locations may be selected again on the basis of the entropy method by ensuring maximum gain of information.

CASE STUDY

An illustration of the entropy based optimization procedure explained in the previous section was applied to an existing rainfall measuring station network. The stations were ranked based on their priority in order to be retained in the network. The changes in the values of correlation and redundancy while adding a new station to the network was also determined. Theses values would help a designer in fixing the optimum number of stations. The details about the study area and the procedure followed in determining the ranks or priorities of different stations in the network are explained.

Network Description

The network considered for the present study is a rainfall measuring network which comprises of about 22 stations extending from Armagh (2878E 3458N) to Cambridge (5435E 2606N). The data required for the study, monthly rainfall data in millimeters was obtained from www.metoffice.gov.uk, United Kingdom Government site for meteorological data. Among the 22 stations two of the stations were non-operative from 2002. So the rest 20 existing stations were considered for the analysis. The data for a time period of about 40 years from 1965–2005 were taken. The positions of the stations are represented in Figure 1 along with station codes.

Analysis of the Problem

The analysis of the problem was carried out as per the methodology explained above. The data series is assumed to follow two types of distributions: normal and log normal. The steps involved in the analysis with normal distribution for the data series is explained.

As the first step the covariance matrix was prepared using the rainfall data and the marginal entropies for each station were determined. Based on the marginal entropy values given in Table 1, the station with rank 1 was fixed as Station 11, which has the highest value for marginal entropy.

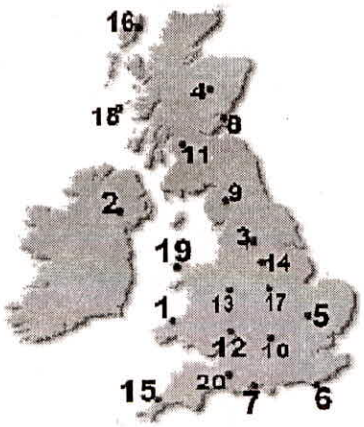

Fig. 1: Rainfall Stations in the Network
(1. Aberporth 2. Armagh 3. Bradford 4. Braemar 5. Cambridge 6. Eastbourne 7. Hum 8. Leuchars 9. NewtonRigg 10. Oxford 11. Paisley 12. Ross-on-wye 13. Shawbury 14. Sheffield 15. St.Mawgan 16. Stornoway 17. Suttonbonington 18. Tiree 19. Valley 20. Yeovilton)

Table 1: Marginal Entropy Values (Normal distribution)

Station ID	*Marginal Entropy*
1	5.114
2	4.866
3	5.113
4	5.148
5	4.725
6	5.248
7	5.226
8	4.884
9	5.198
10	5.074
11	**5.436**
12	4.981
13	4.766
14	5.103
15	5.263
16	5.359
17	4.771
18	5.360
19	5.038
20	4.977

Next the selected node 11 was coupled with all other stations to determine the combination which gives the least transinformation, thus forming the second priority station. For each combination the transinformation is calculated based on the values of conditional entropy determined using Eqn. 6 and Joint entropy determined using Eqn. 8 with $M = 2$ and co-varaiance matrix obtained corresponding to the stations coupled. Table 2 shows the calculations involved in the determination of the second priority station. From the Table, the station 5, which gives the least trans-information value when coupled with station 11 forms the next priority station. The process of coupling the ranked stations with other stations is repeated till all the stations in the network are ranked. When each station is being added to the network the values of redundancy and the corresponding correlations are determined based on Eqns. 13 and 11, and the stations are ranked based on these values. The procedure is repeated for the rain gauge network by considering the monthly rainfall data from the twenty stations assumed to follow a log-normal distribution.

Table 2: Entropy Parameters for the Selection of Second Priority Station

Station Number	*Joint Entropy*	*Conditional Entropy*	*Transinformation*
1	10.379	5.265	0.17033
2	10.061	5.195	0.24072
3	10.393	5.280	0.15588
4	10.296	5.148	0.28781
5	**10.139**	**5.414**	**0.02176**
6	10.621	5.362	0.07420
7	10.556	5.330	0.10581
8	10.157	5.273	0.16229
9	10.160	4.961	0.47439
10	10.479	5.404	0.03156
12	10.337	5.356	0.07932
13	10.147	5.381	0.05486
14	10.454	5.352	0.08403
15	10.545	5.281	0.15474
16	10.351	4.991	0.44462
17	10.171	5.400	0.03567
18	10.231	4.871	0.56440
19	10.280	5.241	0.19433
20	10.347	5.370	0.06616

RESULTS AND DISCUSSION

The results of the analysis with the monthly rainfall values of the gauging stations assumed to follow a Normal distribution are given in Table 3. Based on the transinformation values given in Table 3, the stations are ranked as in Column 2. The correlation co-efficient when each new station is added to the network is given in column 6 of Table. From the Table one can identify which are the stations to be retained and removed for a particular level of redundancy. Moreover the designer can specify the amount of redundancy and select the combination for a particular value of redundancy.

The results of the analysis with the monthly station rainfall values assumed to follow a log normal distribution is given in Table 4. The details of marginal entropy, Joint entropy and Conditional entropy values, and the ranking of stations for the case are given in Table 4.

Table 3: Ranking of Stations (Normal Distribution)

Stn. Rank	Stn. ID	Marginal Entropy	Joint Entropy	Conditional Entropy	Correlation Coefficient R(%)	Transinformation
1	11	5.436	–	–	–	–
2	5	4.725	10.139	5.414	20.64	.022
3	8	4.884	14.782	9.898	61.85	.241
4	10	5.074	19.592	14.518	64.05	.264
5	6	5.248	24.498	19.250	70.41	.342
6	14	5.103	29.234	24.131	72.08	.367
7	2	4.866	33.670	28.804	75.95	.430
8	19	5.038	38.216	33.177	79.16	.492
9	4	5.148	42.860	37.713	79.64	.503
10	16	5.359	47.673	42.313	81.57	.547
11	20	4.977	52.096	47.119	81.84	.554
12	13	4.766	56.273	51.507	83.18	.589
13	9	5.198	60.776	55.578	86.66	.695
14	15	5.263	65.343	60.079	86.71	.697
15	17	4.771	69.401	64.631	87.14	.712
16	12	4.981	73.633	68.653	88.10	.748
17	1	5.114	77.956	72.843	89.12	.790
18	18	5.360	82.515	77.155	89.37	.802
19	7	5.226	86.901	81.674	90.21	.840
20	3	5.113	91.080	85.974	91.84	.927

Table 4: Ranking of Stations (Log Normal Distribution)

Stn. Rank	Stn. ID	Marginal Entropy	Joint Entropy	Conditional Entropy	Correlation Coefficient R(%)	Transinformation
1	6	1.221	–	–	–	–
2	16	0.890	10.13883	5.41399	22.39	0.257
3	8	1.006	14.78166	9.89771	53.17	0.166
4	13	0.914	19.59205	14.51766	69.68	0.332
5	2	0.849	24.49798	19.24975	74.46	0.404
6	5	1.052	29.23416	24.13146	74.54	0.405
7	20	1.085	33.66967	28.80420	79.41	0.498
8	4	0.887	38.21556	33.17718	81.15	0.537
9	14	0.994	42.86042	37.71271	82.06	0.560
10	19	0.952	47.67261	42.31318	83.92	0.609
11	9	0.987	52.09589	47.11854	85.14	0.645
12	18	0.909	56.27301	51.50711	86.04	0.674
13	15	1.045	60.77620	55.57799	87.43	0.723
14	17	0.954	65.34284	60.07938	88.53	0.766
15	12	1.087	69.40099	64.63056	89.13	0.791
16	1	1.028	73.63309	68.65251	89.30	0.799
17	11	1.011	77.95645	72.84273	89.56	0.810
18	10	1.065	82.51457	77.15486	91.00	0.881
19	7	1.196	86.90069	81.67428	91.57	0.911
20	3	0.971	91.08649	85.97364	92.00	0.937

The ranking of stations obtained on the basis of redundancy values with station rainfall data station values assumed to follow normal and log normal distributions are given in column 2 of Tables 3 and 4 respectively. From the tables, it can be seen that there is variation in the ranks of stations as the distribution varies. But if one consider to retain some specified number of stations (say ten) in the network, it can be seen that with few exceptions the selected stations in both the distributions remain same as indicated in Figure 2.

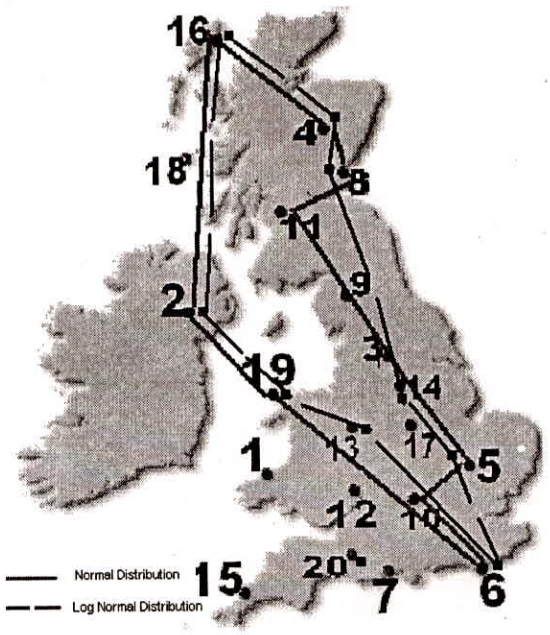

Fig. 2: Optimized rain gauge station network for the two types of distributions

This may be an indication that the importance of distribution type becomes less significant if we are considering a network having large number of stations. In an earlier study (Sarlak and Sorman, 2006), considering a network of four stations in a stream flow gauging network, and it was found that the ranks of stations vary totally, with the type of distribution assumed for the station data series, and the reason for the same was stated as the lesser number of stations in the network. In the present study also the ranks of stations based on the redundancy values are found to vary and depends on the assumed distributions of the data, confirming to the conclusions derived from the study done for stream flow gauging networks (Sarlak and Sorman, 2006).

CONCLUSIONS

The entropy based ranking method has been applied to a rainfall station network. The method helps to develop the information gathering system of a region as it helps to prioritize the stations and avoid the redundant stations thereby giving economical benefits. The analysis was done with two types of distributions namely normal and log normal distribution, to study the effect of distribution in the ranking process. The procedure described helps to assess a network with respect to the existing monitoring sites. If new stations are to be added to the network their locations may be selected on the basis of entropy method by assuring the maximum gain of information.

REFERENCES

Kapur, J.N. and Kesavan, J.K. (1992). *Entropy optimization principles with applications*, Academic Press Inc.

Ozkul, S. Harmaucioglu, N.B. and Singh, P.V. (2000), "Entropy based assessment of water quality monitoring networks", *Journal of Hydrologic Engineering*, ASCE, 5(1), 90–99.

Sarlak, N. and Sorman, U.A. (2006). "Evaluation and selection of streamflow network stations using entropy methods", *Turkish Journal of Eng. Env. Sci.*, 30, 91–100.

Singh, P.V. (2000). "The entropy theory as a tool for modelling and decision making in Environmental and Water Resources", *Water SA*, 26(1), 1–11.

UK Government Site for Meteorological Data, http://www.metoffice.gov.uk.